

# Reply to: “A response to some unwarranted criticisms of single-grain dating” by J.K. Feathers

**Thomsen, Kristina Jørkov; Murray, Andrew Sean; Buylaert, Jan-Pieter; Jain, Mayank; Helt-Hansen, Jakob; Aubry, Thierry; Guerin, Guillaume**

*Published in:*
Quaternary Geochronology

*Link to article, DOI:*
10.1016/j.quageo.2016.10.007

*Publication date:*
2017

*Document Version*
Peer reviewed version

# REPLY TO: "A RESPONSE TO SOME UNWARRANTED CRITICISMS OF SINGLE-GRAIN DATING" BY J.K. FEATHERS

Kristina J. Thomsen[a], Andrew S. Murray[b], Jan-Pieter Buylaert[a,b], Mayank Jain[a], Jakob H. Hansen[a], Thierry Aubry[c]; Guillaume Guérin[d]

[a] Center for Nuclear Technologies, Technical University of Denmark, DTU Risø Campus, DK-4000 Roskilde, Denmark
[b] Nordic Laboratory for Luminescence Dating, Department of Geoscience, Aarhus University, DTU Nutech, Risø Campus, DK-4000 Roskilde, Denmark
[c] Côa Parque, Fundação para a Salvaguarda e Valorização do Vale do Côa, Rua do Museu, 5150-610 Vila Nova de Foz Côa, Portugal
[d] Institut de Recherche sur les Archéomatériaux, UMR 5060 CNRS - Université Bordeaux Montaigne, Centre de Recherche en Physique Appliquée à l'Archéologie (CRP2A), Maison de l'archéologie, 33607 Pessac cedex

In the note "A response to some unwarranted criticisms of single-grain dating" Feathers raises many issues with both the approach and the conclusions of Thomsen *et al.* (2016). After careful consideration, we find we disagree with Feather's analysis and conclusions, and stand by the original conclusions of Thomsen *et al.* (2016). We reiterate that, for these samples, the multi-grain measurements are demonstrably in better agreement with the independent age control than are the standard single-grain measurements.

In our view, Feathers' most important criticisms are that the $^{14}$C age control is reported incorrectly and that Thomsen *et al.* (2016) cannot conclude that standard single-grain methods are in poorer agreement with the independent age control than the multi-grain methods. We acknowledge the presence of a minor presentation error in Figure 3 of Thomsen *et al.* (2016), but we demonstrate that this detail has no bearing on the conclusions of Thomsen *et al.* (2016).

We respond below in detail to the main issues raised by Feathers. We have retained his structure for ease of cross-comparison.

## 1 INTRODUCTION

In the introduction to his comment, Feathers "*begin[s] by pointing out that the literature contains many instances of single-grain OSL ages in good agreement with independent evidence, even for well-bleached and unmixed samples".* Not only are the references given by Feathers on the subject rather incomplete (they exclude in particular a recent review on the subject, see below), but most are irrelevant or misinterpreted. Feathers himself acknowledges: *"The main purpose of this study [Thomsen et al., 2016] is to investigate the accuracy of single-grain OSL dating beyond 20 ka using four samples with associated $^{14}$C ages from the Bordes-Fitte rockshelter*". Given this perspective, 2 out of 5 references on which Feathers' argument is based can be dismissed: in Feathers *et al.* (2010), the samples to which he refers, for which OSL and $^{14}$C agree, are <20 ka, while in Arnold and Demuro (2015) the single grain luminescence ages are TT-OSL ages and not OSL ages. Looking in detail at the third reference (Demuro *et al.*, 2013) used by Feathers to support his statement reveals that (i) there is no direct age control for the samples dated by OSL (the only chronological information is bracketing tephra layers); excluding the poorly-defined, intermittent tephra SCt-A, the only way to test the hypothesis that SG-OSL ages underestimate independent ages is to look at two samples (53A and OQC10), which should be older than tephra SCt-K dated to 77 ± 8 ka; (ii) 59 and 42 dose estimates from small aliquots (not single grains, each aliquot contained ~1 to 18 grains) were

incorporated in the $D_e$ distributions; these were significantly over-dispersed (OD: 27-41%). As a result of these difficulties, there are large uncertainties on both the OSL and tephra ages, and the ratio of pseudo single-grain OSL to overlying tephra ages is 0.90 ± 0.10. We conclude that the Demuro *et al.* (2013) study is insufficiently precise to be used to distinguish between Feathers' claimed agreement and the degree of underestimation observed by Thomsen *et al.* (2016).

The last two papers cited by Feathers in support of his argument (Douka *et al.*, 2014 and Jacobs *et al.*, 2015), indeed provide single-grain CAM-based ages in agreement with independent chronologies. However, it should be noted that in both cases, multi-grain luminescence ages (OSL ages in Douka *et al.*, 2014; MET-pIRIR ages in Jacobs *et al.*, 2015) also agree with independently-derived ages.

Furthermore, Feathers does not discuss a recent study by Guérin *et al.* (2015a) in which multi-grain OSL ages and single-grain CAM-based ages were compared with reference ages. This study included 19 samples (including the 4 from Thomsen *et al.*, 2016) from various sites measured in two different laboratories by five different users; the average multi-grain OSL to reference age ratio is 0.97 ± 0.03 (n=12), while the average CAM-based single-grain OSL to reference age ratio is 0.90 ± 0.02. Their Fig. 5 clearly shows a systematic increase of age underestimation with CAM-based single-grain OSL as a function of independent age. In other words, the results reported by Thomsen *et al.* (2016) are not unusual; single-grain CAM ages do repeatedly underestimate independent age control in cases where multi-grain ages do not.

## 2 INDEPENDENT AGE COMPARISONS

Based on the agreement between multi-grain feldspar and multi-grain quartz, Thomsen *et al*. (2016) were able to conclude that it is extremely unlikely that the quartz is poorly bleached and therefore overestimates the deposition age. They then pointed out that the multi-grain quartz doses were consistent with the dose expected from the known dose rates and the $^{14}C$ ages. Based on this, they concluded that there was no evidence to question the reliability of the multi-grain quartz doses and then proceeded to test single grain doses against these multi-grain doses. This has the advantage of being a more precise test than comparing with another dating technique because the uncertainty associated with the dose rate is removed from the comparison. Feathers rightly argues that Thomsen *et al*. (2016) could have directly compared both multi-grain and single-grain results with the $^{14}C$ age control. Unfortunately, this would be at the expense both of these increased uncertainties and at the cost of reducing the comparison from four pairs to three pairs; OSL sample 092204 cannot be compared because the independent age control is too wide (~22 to >36 ka) to be useful.

In his section 2, Feathers first criticises Thomsen *et al's*. (2016) use of the $^{14}C$ upper limit for OSL samples 092203 and -04. He appears to misunderstand the arguments behind the discussion of the older end of the age control on these two samples. Since they address exactly this issue in the original text (see p. 79 and 80) and qualify in detail their use of $^{14}C$ sample Beta-234193, his comment seems to be redundant. More importantly, his criticism of the use of these $^{14}C$ data is actually irrelevant to the debate on the accuracy or otherwise of our CAM single grain doses. Thomsen *et al*. (2016) clearly state (p. 80) that OSL sample 092203 was taken adjacent to and at the same depth as $^{14}C$ sample OXA-22315 (23322-22615 cal yr, CI 95%). They thus have a closely associated age control for OSL sample 092203. Their multi-grain OSL dose is completely consistent with the expected dose based on this $^{14}C$ age, whereas their CAM single-grain dose (using

standard rejection criteria) is inconsistently low. These statements are independent of any uncertainty on the upper limit to age of the unit and Feathers does not seem to disagree with this (see his Table 1).

On the other hand, Feathers is correct to point out that Thomsen *et al*. (2016) were inconsistent in the way they showed the predicted doses based on $^{14}$C shown in their Figure 3. There they used the 95% CI interval for the $^{14}$C and the 68% CI on dose rates; these, of course, should both have been at the same CI (68% is conventional in OSL dating, but the choice is arbitrary). The effect of this small error is to slightly reduce the uncertainties in the predicted doses, but it does not change the statistical outcome of a comparison of predicted to measured doses.

However, Feathers' claim that "*the single-grain results do not disagree with the radiocarbon controls any more than the multi-grain results do*", is factually incorrect and the P-values, obtained from a two-sample $\chi^2$ homogeneity test and quoted in his Table 1, have been incorrectly calculated. This error appears to arise both because of the effect of using rounded data and incorrect estimation of the uncertainty of the $^{14}$C predicted quartz doses. For example, the $^{14}$C sample OxA-22315 has an age range of 23,322-22,615 calibrated years before 2009 (CI 95%), which corresponds to 22,969±354 cal yr (CI 68%, where the uncertainty is derived by dividing the range by 3.92, based on the quantile function of a Gaussian distribution; that is assuming for the purpose of comparison with another dating method that the $^{14}$C uncertainty is normally distributed). The dose rate for the corresponding OSL sample 092203 is 3.235±0.148 Gy/ka, which gives a predicted $^{14}$C dose of 74.3±3.6 Gy. The uncertainty quoted by Feathers is ±8 (at CI 95%) which gives for comparison ±4.1 (= 8/1.96) at CI 68%, *i.e.* 14% larger than our value. Feathers' reported uncertainties for the $^{14}$C predicted quartz doses are, on average, 23±5% (n=4) larger than our values based on original data and this leads to a significant difference in the reported P-values. In Table 1, we report our calculated P-values for multi-grain and single-grain (CAM) quartz doses obtained using standard rejection criteria (*i.e.* not making use of the $D_0$ or FR criteria). We agree with Feathers that all multi-grain quartz doses are consistent with the $^{14}$C predicted doses, but of the three relevant single-grain CAM quartz doses, only sample 092201 (with a P-value of 0.07) cannot be argued to be statistically different from the $^{14}$C age control at the standard threshold of 0.05. However, it is important to note that the average ratio (unweighted) between the single-grain CAM dose and that predicted from $^{14}$C is 0.877±0.015 (n=3) and the weighted mean ratio is 0.88±0.03; neither of these are consistent with unity, whereas the same ratios for multi-grain quartz doses are 1.00±0.03 (unweighted and weighted). Thus, we do not agree with Feathers statement that "*the single-grain results do not disagree with the radiocarbon controls any more than the multi-grain results do*".

Although the $^{14}$C age control for sample 092204 is too broad to be useful, given our confidence in the multi-grain doses (see above), we can now compare the single-grain to multi-grain doses for all four samples (as is done in Thomsen *et al.*, 2016). The average unweighted ratio is 0.866±0.019 (n=4) and the weighted ratio is 0.865±0.018 (n=4). These are indistinguishable from the ratios of doses from single-grain and $^{14}$C discussed above and thus supports the choice of directly comparing the single-grain results with multi-grain results.

## 3 THE IMPORTANCE OF SAR REJECTION CRITERIA

Feathers disagrees with the suggestion "*that single-grain dating studies [should] document the effect on dose and dispersion of applying routine rejection criteria*" apparently because he is convinced that one should reject what are, in his view, "*inaccurate grain types*" even if such inaccuracy cannot be demonstrated. Thomsen *et al.* (2016) show that applying "standard" rejection criteria (*i.e.* grains are rejected if the recycling and IR depletion ratios are inconsistent with unity at two standard deviations and if the absolute measured recuperation dose is larger than and inconsistent with 1 Gy) does not change either the CAM dose or the over-dispersion significantly and they thus conclude that these criteria demonstrably do not contribute to providing a more accurate or precise data set. There is therefore no benefit in applying these criteria to these samples. Similar conclusions have been reached by e.g. Geach *et al.* (2015), Guérin *et al.* (2015b), Hansen *et al.* (2015), Kristensen *et al.* (2015), Zhao *et al.*, (2015) based on published data. Indeed Feathers himself (Feathers *et al.*, 2010, p. 418) has commented that the acceptance threshold for recycling can be relaxed, (thereby accepting grains that are "*known to violate the assumption of the method*") without significantly affecting results and that this allowed a larger sample size.

To our knowledge, there are only two papers that provide published numerical support for the employment of these criteria: Jacobs *et al.*, (2006), identified on p. 93 in Thomsen *et al.* (2016) and Doerschner *et al*. (2016), which was published after Thomsen *et al.* (2016). Feathers appears to support the hypothesis that these standard rejection criteria must provide a significant improvement in accuracy and precision, based in particular on a recent article published by Jacobs *et al*. (2015). However, in this article, despite a lengthy discussion of the different types of grains rejected for various reasons, no numerical comparison is made of the parameters describing the $D_e$ distributions (*e.g.* CAM and OD) before and after rejecting the grains (apart from the effect of rejecting grains based on their FR). In the great majority of cases where the hypothesis that standard rejection criteria must provide an improvement in accuracy and precision has been tested, it has been shown to be untenable. Thus, we stand by the recommendation of Thomsen *et al*. (2016) that single-grain studies (and indeed all OSL dating studies) should document the effect on dose and dispersion of applying routine rejection criteria, and indeed we are surprised at the disagreement caused by such a simple suggestion.

Feathers correctly points out that in testing the standard rejection criteria Thomsen *et al*. (2016) include the widely adopted approach of arbitrarily rejecting grains with natural signal lying at or above the saturation level of the laboratory dose response curve. This approach has indeed been routinely applied in single-grain dating applications for many years; mainly because there has been no other obvious way of processing these data. However, Thomsen *et al.* (2016) point out that such a process will inevitably lead to a bias in dose estimation, as can readily be realised by considering the effect of applying this criterion to a conceptual sample in saturation (where every grain is well-behaved). Because of the inherent dispersion in signal measurement, ~50% of the sensitivity–corrected natural signals will lie above the laboratory dose response curve (with no intercept), and ~50% will provide a finite estimate of dose. If the non-intercepting signals are rejected, then the remaining signals will appear to give a finite dose in the sample, rather than the non-finite dose that should be measured. The same bias will apply to a lesser degree to samples approaching saturation. As Feathers points out, Thomsen *et al.* (2016) do single this approach out for special attention; they do this because, from first principles, it will result in a bias in dose distributions, especially those derived from single-grain measurements.

However, Feathers speculates that sensitivity corrected natural signals ($L_n/T_n$) that do not intercept the dose response curve have arisen for other undocumented reasons and that therefore this bias is not inherent – of course this may be true if these undocumented reasons only produce $L_n/T_n$ signals which lie above the laboratory dose response curve. To our knowledge there is no evidence for this, and one can as readily speculate that any "*experimental artefacts*" or "*inherently unreliable grain response*" would produce lower than expected results as well as greater than expected. Feathers then suggests that: "*The high proportion of saturated and nonintersecting grains … and the improvement in the dose recovery ratio when those grains are removed (Figure 7) suggests these are not just a matter of genuinely older grains, but rather an experimental artefact".* Unfortunately, this comment demonstrates a misunderstanding of the rejection criterion. In Figure 7, Thomsen *et al*. (2016) do not reject grains based on whether or not their natural signal intersects the corresponding dose response curve, but rather based on their $D_0$ value. The point here is that for a grain to be accepted into an analysis it must be capable of recording the dose of interest. For instance, it would be impossible to measure a dose of 100 Gy with a dosimeter with a single saturating exponential dose response curve with $D_0$ = 10 Gy. But if this dosimeter consisted of many hundreds of sensitive quartz grains (all of the same $D_0$) it would of course be inevitable that ~50% would give a finite, measurable dose, purely because of dispersion in $L_n/T_n$. However, these finite dose estimates would all be underestimates and part of a population of grains that should be rejected in its entirety, as all these grains are unable to record the dose of interest.

Feathers is further confused between the rejection of grains because they are saturated, or do not intersect the dose response curve, and rejection of grains because of an insufficiently $D_0$. In the above quotation, Feathers suggests that the dose recovery ratio is improved by the simple removal of saturating and non-intersecting grains. First of all, one cannot calculate a bounded dose recovery ratio without discarding all unbounded dose estimates (*i.e.* all estimates for which $L_n/T_n$ lies within one standard deviation of laboratory saturation, or above saturation). Thus, Thomsen *et al*. (2016) did not, as Feathers suggests, obtain an improvement in dose recovery by discarding saturated grains. This improvement in dose estimates was in fact obtained by the implementation of the $D_0$ criterion, *i.e.* they rejected all grains for which the $D_0$ was less than the (known) given dose, thus building their dose distributions only from those grains with a $D_0$ sufficiently large, that the average $L_n/T_n$ of the given dose lay at or below 68% of the relevant saturation light level. Finally, it is entirely to be expected that the proportion of grains in saturation increases with given dose in dose recovery experiments. This simply reflects the fact that more and more grains have insufficient $D_0$ values to measure the $L_n/T_n$.

## 4 INAPPROPRIATE USE AND INTERPRETATION OF STATISTICAL AGE MODELS

In OSL single-grain dating there is a range of models routinely applied to extract the burial dose from single-grain dose distributions. Thomsen *et al.* (2016) show that, for instance, for sample 092201 these different models can lead to apparent burial ages ranging between 23±2 and 53±4 ka, illustrating that the choice of model is extremely important. The justification for applying each model is clearly given in the text and in every case this justification has been used in previously published OSL studies. Feathers does not seem to acknowledge that Thomsen *et al.* (2016) are not advocating the use or otherwise of any of these models.

They simply test whether or not models that others have used in seemingly relevant contexts provide accurate ages for our samples.

Feathers starts by commenting that Thomsen *et al*. (2016) wrongly claim that an over-dispersion (OD) threshold of 20% has been recommended in the literature to identify the possible existence of partial bleaching, despite the fact that they give references supporting exactly this claim. Olley *et al.* (2004b) explicitly state that: "*In cases where the data over-dispersion suggests partial or heterogeneous bleaching of the OSL signal (rd>20% for single grains), the minimum age model should be used to estimate the burial dose from the lowest De population. For samples that appear to have been well bleached at the time of deposition (rd < 20% for single grains), the central age model should be used to calculate the burial dose."* Jacobs *et al.* (2015) also state explicitly that one of their three criteria for deciding which model to apply is an OD of >20%. Arnold and Roberts (2009) use the value of 20% as *"…a useful approximation for the common dispersion parameter, s, of the FMM…*" clearly implying that this is perceived as an upper limit to over-dispersion for an unmixed (and presumably well-bleached) dose distribution - although they do go on to advocate the use of independent statistical criteria such the BIC score to refine this choice.

Feathers then criticizes Thomsen *et al.* (2016) for applying models blindly. He suggests that models (in this case particularly the decision tree of Bailey and Arnold, 2006) that were developed for fluvial systems are not relevant to their colluvial sediments. *(*Later he also states: *"They choose to apply statistical approaches developed in non-comparable depositional / dosing contexts or developed for completely unrelated samples".*) However, this decision tree has been applied to many different depositional environments including fluvial, glacio-fluvial, colluvial, alluvial, coastal and aeolian (e.g. Fuchs and Owen, 2008; Delong and Arnold, 2007; Fattahi *et al*., 2010; Costas *et al.*, 2012; Stone *et al.*, 2010; Gaar and Preusser, 2012), to determine whether minimum dose modelling is appropriate. Applying this decision process to single-grain dose distributions of colluvial origin is thus fully justified by the documented use in the literature. The outcome clearly indicates that all the natural dose distributions given by Thomsen *et al*. (2016) should be analysed using one of various minimum age models. However, when they do this, they obtain significant underestimates of the predicted dose.

Feathers is also concerned by the testing of models on laboratory-generated dose distributions where the answer is already known. He "*finds it particularly odd*" that Thomsen *et al.* (2016) apply the decision tree to dose recovery data and states that "*Misapplications of published procedures do not demonstrate their poor suitability. Rather they highlight the obvious problems that can arise from carelessly applying statistical analyses and age modelling in inappropriate contexts*." We find this statement very surprising. Any process or model that is unable to reproduce the expected value in controlled experiments, where the outcome is known, is of limited value. 'Benchmarking' under different boundary conditions is common scientific practice in numerical modelling. For instance, if a model finds multiple dose components in a laboratory well-bleached sample, it is likely that it will incorrectly predict the same in a well-bleached natural sample. The fact that the decision tree only indicates a well-bleached dose distribution in 25% of the dose recovery experiments is a strong indication that it is unable to reliably distinguish between well-bleached and partially-bleached dose distributions in natural samples. This is of considerable concern, because this decision tree is one of the few non-subjective methods proposed for choosing the most appropriate model with which to analyse single-grain dose distributions. The failure of the decision tree to identify the

appropriate model in dose recovery experiment indicates that it is also unreliable for the identification of appropriate models when applied to natural samples.

However in real dating scenarios the selection of an appropriate model is, in any case, largely a subjective decision. For instance, Jacobs *et al.* (2015) state that they use three criteria for the selection of the most applicable model: (1) OD > 20%, (2) Visual examination of radial plots (presumably constructed using only known analytical uncertainties), and (3) their knowledge of the site and sample context. Only the first of these criteria is non-subjective (although arbitrary) and that is effectively the 20% criterion that Feathers denies is used in the literature. Various authors have addressed the problem of relying on visual examination of radial plots (e.g. Thomsen *et al*., 2016; Reimann et al., 2012; Guérin et al., 2016). In our view, the inevitable subjectivity of (2) and (3) above is a cause of considerable concern and it is very important to test whether such subjective decisions might influence the outcomes.

Feathers also claims that the application "*of the FMM is similarly unconventional*", because Thomsen *et al.* (2016) choose to apply it both to the natural dose distributions and to controlled laboratory experiments (*i.e.* to dose recovery experiments). When applying the FMM to the dose recovery distributions, the "additional uncertainty" is determined by optimising the BIC and the llik scores (Galbraith, 2005) and this gives values of between 9 and 20%. Feathers states that "*given that the over-dispersion from their single population dose recovery tests is 14-29% (their Table 4), the optimized additional uncertainty used with the FMM should be at least that much*". This is incorrect - the reported over-dispersion values given in Table 4 of Thomsen *et al.* (2016) range between 7±1 and 29±3%; clearly there are dose recovery ODs smaller than the 9% OD derived from optimizing the FMM input parameters. In any case, when the optimization of the BIC and llik scores is used in the literature, a fixed predefined additional uncertainty interval is normally used, regardless of the over-dispersion value(s) determined in dose recovery experiments (e.g. Jacobs *et al.* 2012). It appears that others use the FMM in a similarly "unconventional" manner.

## 5 FAST RATIO

Feathers appears to be concerned that Thomsen *et al.* (2016) find the use of the fast ratio FR as a selection criterion expensive in terms of data rejection because of variation in effective stimulation power. Note that both Thomsen *et al.* (2016) and Feathers seem to agree that the FR does discriminate in favour of grains giving more accurate estimates of dose.

Feathers suggests that one way to reduce the number of rejected grains would be to increase the FR cut-off incrementally until the average dose formed a plateau; further, he speculates that a lower appropriate threshold of *e.g.* FR>2 might help the situation. On p. 91 Thomsen *et al.* (2016) state: "*However, although only choosing the grains with the largest FR values is very expensive in terms of data reduction, Thomsen et al. (submitted) nevertheless show that the ratio of measured to expected dose increases systematically for these samples, when only grains with FR>4 are included*". While this perhaps could have been phrased more clearly, we have in fact investigated the effect of varying the FR and find that lower values of FR do not give the required systematic improvement in natural dose. Thus, Feathers is incorrect in suggesting that it would be possible to apply a lower FR threshold and thereby accept a larger fraction of grains.

We do not understand Feather's comment that using the FR criterion seems to improve dose recovery reliability. Thomsen *et al.* (2016) make no such claim, nor do they present data that would allow such a

claim. In fact, analysis of their data does not indicate that using the FR criterion improves dose recovery. On the other hand, it is entirely to be expected that application of the FR criterion even in the presence of variation of effective stimulation power will improve the accuracy of equivalent dose estimation. This is because a rapid OSL signal decay (*i.e.* high FR) only occurs for the highest effective stimulation powers and an OSL signal dominated by the fast component. If either the power is low or the OSL signal is not dominated by the fast component the grain is rejected. Thus, since no grains with undesirable components are included in the analysis, the remaining grains are likely to give a more accurate dose estimate.

Feathers states that "*Thomsen et al. (2016) show that low $D_0$ values are less prevalent in their low-dosed dose recovery experiments. They occur more often with high doses.*" This statement is factually incorrect; Thomsen *et al.* (2016) do not present such data nor would we expect to be able to: the $D_0$ value is a characteristic of the luminescence response of the grain and is independent of the $D_e$. In Table 4 of Thomsen *et al.* (2016), the number of grains in saturation in the various dose recovery experiments is quoted; it is shown that, not surprisingly, the smaller the given dose the fewer grains are rejected because of saturation. However, we only provide the distribution of $D_0$ values for a single dose recovery experiment (Figure 7). Thus, Feathers again appears to confuse the presence of saturating grains with the values of $D_0$ (see section 3 above).

## 6 STATISTICS

Feathers claims that Thomsen *et al.* (2016) "*ignore basic statistical principles and draw unjustified conclusions*" and he seems to base this statement on two references to the original manuscript. He states that "*Thomsen et al. (2016) argue that for "multi-grain dose distributions it is clear there is no advantage in deriving a CAM dose in preference to an average dose" (p. 85), by which they mean an unweighted arithmetic mean. Later, in comparing single- to multi-grain dose, they imply that the arithmetic mean is better for single-grain distributions as well, arguing "that the uncertainties used for weighting in CAM are inappropriate" (p. 88).*" First of all, Thomsen *et al.* (2016) do not argue that for multi-grain dose distributions there is no advantage in deriving a CAM dose. They observe this to be the case; it is quite simply an experimental result. With regard to the second quotation, Thomsen *et al.* (2016) state that the observation that the CAM single-grain doses significantly underestimate the multi-grain doses (and the doses predicted from $^{14}C$, see above) seems "*to imply that the uncertainties used for weighting in CAM are inappropriate*". However, in the very next sentence, they then go on to show that the same underestimation occurs if they do not use weighting (*i.e.* calculate an arithmetic average), thus negating the implication. So Thomsen *et al.* (2016) do not imply that an arithmetic mean is a more accurate estimator for single-grain distributions, as Feathers suggests. Feathers also neglects to refer to the conclusion of the discussion section in which Thomsen *et al.* (2016) point out that they do not observe an improvement in the arithmetic average of the measured-to-expected natural dose ratio when employing the additional rejection criteria ($D_0$ and FR) – both averages are just inconsistent with unity. However, they go on to say that "*a significant improvement in accuracy is observed for CAM*; *the measured-to-expected ratio is increased from 0.87 ± 0.02 to 1.04 ± 0.02…*" Thus, their observation is that under these circumstances the CAM results are accurate; presumably implying that the weighting is valid. Thus, at no point do Thomsen *et al.* (2016) "*ignore basic statistical principles and draw unjustified conclusions*". They simply draw conclusions from experimental observations and Feathers misrepresents these conclusions.

Later in this section, Feathers takes issue with the reference to Guérin *et al.* (2013) and states that Galbraith (2015a) has dismissed their arguments. However, with reference to Guérin *et al.* (2013), Galbraith (2015b) clearly states that "*The above argument … supports their suggestion that, when all grains have the same true age, then this age may be estimated using an average or central age, regardless of how the individual dose rates vary",* although he does go on to say that this does not necessarily imply that the average is the best method. Nevertheless, we regard this as an ongoing discussion; it is certainly not as clear cut as Feathers suggests.

In passing, Feathers also defends here the use of the FMM to investigate beta dose heterogeneity despite the observation of Thomsen *et al.* (2016) that "phantom components" can be observed even in dose recovery experiments. But of course if, under laboratory conditions, the FMM identifies discrete components in dose distributions where these are known not to exist, then it must be assumed that the FMM may also incorrectly identify the existence of components in natural dose distributions. Whether these "phantom components" are assumed to arise from beta dose heterogeneity or mixing is irrelevant if they do not exist.

## 7 How can the multi-grain ages be reliable if the single-grain ages are questioned?

Feathers claims that "*there is a fairly striking omission in Thomsen et al.'s (2016) argument that the multi-grain OSL ages are more reliable than the single-grain OSL ages at Bordes-Fitte, and this is only briefly acknowledged in their conclusion*". He goes on to ask: "*How is it then possible that the multi-grain OSL measurements, which included all these aberrant grain populations … gave "the most accurate ages"…*?" Thomsen *et al.* (2016) fully acknowledge the observation that their experimental results are inconsistent with expectations; they write on page 95: "*Given the very large fraction of single-grains that must be rejected to provide accurate single-grain dose estimates, it is of course surprising that multi-grain dose estimates (based on the sum of signals from acceptable as well as unacceptable individual grains) provide accurate dose estimates without any further data selection*".

It appears to us that Feathers uses his expectation of the unreliability of multi-grain results to implicitly question the validity of the experimental observation. Thomsen *et al.* (2016) state that "*had these samples been analysed in the absence of other age control, the application of standard single-grain methods would have led to significant misinterpretations of results and a corresponding inaccuracy in ages*". Again, it is an experimental observation that multi-grain results are in better agreement with the independent age control than are single-grain results (by following the various standard procedures used in the literature, all attempts to extract single grain ages were substantially inaccurate, by up to 50%). It is true that when the FR and $D_0$ criteria were applied and the doses estimated using CAM, the single-grain ages became acceptably accurate. But note that for samples with no independent age control this would require advanced knowledge that the distributions were unmixed and well-bleached. The objective methods of examining these dose distributions (*i.e.* the OD 20% threshold and the decision tree of Bailey and Arnold, 2006) identify these samples as poorly bleached. And in our view the subjective approaches (e.g. visual inspection of radial plots and knowledge of deposition environment) would probably have concluded that the samples were either poorly bleached or mixed (since these samples are from colluvial deposits with

over-dispersions >30%). Thus, we cannot agree with Feathers' argument that the data of Thomsen *et al.* (2016) in any way contradict their position.

Finally, Feathers criticises our dose recovery data as being inaccurate ("*Such dose recovery statistics do not appear to provide overly strong support for the suitability of the single-grain or multi-grain SAR procedure adopted in this study*") and argues that the use by Thomsen *et al.* (2016) of a dose recovery within 10% of unity as being satisfactory is inadequate. Here they are essentially being criticised for being too thorough in their dose recovery studies. We now first restrict ourselves to the results that would be obtained in a "standard" dose recovery test (e.g. Demuro *et al.*, 2013) where grains are bleached using blue light at room temperature and subsequently given a dose approximately equal to the natural. Thus, we consider the Thomsen *et al.* (2016) results for 65 Gy for samples 092203 and 092204 and for 100, 110 and 119 Gy for sample 092201. (Unfortunately we do not have a multi-grain beta dose recovery close to the natural dose for sample 092202.) These datasets give average multi-grain dose recovery ratios of: 0.976±0.014 (n=28, sample 092201), 1.05±0.04 (n=8, sample 092203), 0.98±0.02 (n=8, sample 092204). All of these are consistent with unity at two standard deviations; since they are all derived from what is presumably the same type of quartz we can also calculate the grand average of 0.991±0.012 (n=44). Thus, if we restrict ourselves to one of the most commonly used standard dose recovery test, there is no evidence that our multi-grain dose recovery ratios are inconsistent with unity. Similarly, the relevant dose recovery ratios reported for single-grains are: 0.97±0.03 (n=105, 110 Gy, sample 092202) and 0.96±0.02 (n=165, 65 Gy, sample 092201); both of these clearly fulfil Feathers' criterion.

Having pointed out that the relevant dose recovery ratios were in fact satisfactory, we would nevertheless point out that there is little or no evidence that measured dose recovery ratios correlate with equivalent doses. Guérin *et al.* (2015a) saw no significant correlation between observed dose recovery ratios and the accuracy of ages for known-age samples and concluded that they agree with the suggestion of Murray and Wintle (2003) that dose recovery ratios are not necessarily a good indicator of the accuracy of ages obtained with the SAR procedure. In addition, Jacobs and Roberts (2015) show that their equivalent doses measured 7 years apart using two different SAR protocols are indistinguishable, despite the fact that the dose recoveries for three samples obtained using one SAR protocol lay between 0.79±0.04 and 0.92±0.02 whereas dose recoveries for the same three samples obtained using the second protocol lay between 1.02 ±0.02 and 0.97±0.02. In any case, the assertion that a dose recovery must be statistically consistent with unity is misleading. For example, a dose recovery of 0.995 ± 0.001 would, in our view, be perfectly acceptable but not consistent with unity. Even if a deviation in dose recovery ratio from unity led to a corresponding inaccuracy in $D_e$, some systematic error would still be acceptable. This is why Murray and Wintle (2003) did not suggest that a dose recovery ratio must be statistically consistent with unity and presumably why Jacobs and Roberts (2015) required that their dose recoveries should be within 5% of unity.

## 8 CONCLUSION

Feathers lists six issues in his conclusion and claims that Thomsen *et al.* (2016):

1) *Fail to report the independent age control correctly*. We disagree; the exact calibrated and uncalibrated ages are given on page 79. There is a minor issue of presentation of uncertainties on

doses derived from known age. We acknowledge that, in Figure 3 of Thomsen *et al.* (2016), the predicted $^{14}C$ dose interval is shown at a CI of 95% but the uncertainty arising from the dose rate at a CI of 68%. However, we have shown above that correcting this minor graphical error has no bearing on the conclusions of Thomsen *et al.* (2016). The multi-grain results are systematically in better agreement with the $^{14}C$ age control than the single-grain results obtained using standard rejection criteria.

2) *Have missed the point of rejection criteria when they argue that if rejection criteria make no difference to the mean and over-dispersion value of a dose distribution they are of little value*. We disagree. Feathers puts forward the hypothesis that grains that fail these rejection criteria must be inaccurate (despite the fact that he himself has chosen to accept relaxed rejection criteria, because they do not significantly affect his results; Feathers *et al.*, 2010 on p. 418). But this hypothesis is unproven, and in fact the evidence in Thomsen *et al.* (2016) and others suggest that any such effects are undetectable in their samples. Note that Thomsen *et al.* (2016) do not advocate ignoring such rejection criteria but rather suggest that it would be good practice to test and document the effects of these criteria and discuss whether they are of benefit on a case by case basis.

3) *Fail to appreciate how different age models are used in the literature and apply them to samples for which they are clearly inappropriate or are applied in an incorrect manner*. We disagree. In our view, Feathers has completely misunderstood the thrust of this paper. Thomsen *et al.* (2016) show that, not surprisingly, the equivalent doses obtained from single-grain dose distributions are very dependent on the model chosen. But most importantly, they showed that the only non-subjective method of choosing which model to apply is unable to reach the right conclusions. Thus, in the absence of age control, the choice of the appropriate model seems to rest entirely on the subjective opinion of the geochronologist. Feathers may feel happy with this, but in our view this is unacceptable; in general, one cannot decide by looking at a sedimentary section or a radial plot whether a sample is well-bleached or not or whether it was significantly affected by post-depositional processes such as mixing. Thus, it is imperative that we acknowledge the subjectivity of our model results and search for reliable non-subjective decision trees.

4) *Underestimate the potential of the fast ratio FR*. As we discuss above, we think that Feathers has completely misunderstood the position of Thomsen *et al.* (2016). They fully acknowledge that the FR produces an improvement in the accuracy of single-grain doses for these samples. The cut-off point chosen by Thomsen *et al.* (2016) was not arbitrary, as claimed by Feathers, but was in fact the smallest value that gave us the required accuracy. Our conclusion remains that selecting grains based on their FR is very expensive in terms of measurement time; this criterion rejects grains with acceptable luminescence characteristics because these are misidentified as a result of variation in effective stimulation power.

5) *Use unweighted arithmetic means for combining data*. For multi-grain data Thomsen *et al.* (2016) show experimentally that there is no improvement in accuracy in deriving CAM (*i.e.* weighted geometric mean) doses; this is not surprising as statistical uncertainties on large multi-grain aliquots are usually homogeneous and only make a small contribution to the dispersion in

measured $D_e$ values. With respect to single-grain dose distributions, Thomsen *et al.* (2016) explicitly state that they do not observe an improvement when using the arithmetic average (p. 94, last paragraph in "summary and discussion"). Thus, they do indeed calculate arithmetic means, but this is for comparison with other methods; in particular, it should be noted that they do not advocate the use of arithmetic means rather than weighted means in this paper.

6) *Do not explain why the multi-grain results are more accurate than the single-grain results despite the presence of aberrant grains.* We agree. Thomsen *et al.* (2016) do not explain but simply show it to be the case.

At the end of his critique, Feathers concludes that he considers *"… the single-grain results, using all of the rejection criteria, at Bordes-Fitte to be the most reliable..."* We agree that the single-grain results are reliable if ALL rejection criteria are adopted, but we cannot accept that they are any more reliable than the multi-grain data. This is an experimental observation and thus not subject to opinion. But it is very important to realise that rejection criteria, as widely used in the literature by Feathers and others, would not have achieved such a satisfactory result on these samples. Particularly, and despite Feathers' misunderstanding, the $D_0$ criterion is not standard (it was first suggested in Thomsen *et al.*, 2016); the application of the FR criterion to single grains is similarly very recent and has had limited application.

In conclusion, despite a careful reading of Feathers' "*A response to some unwarranted criticism of single-grain dating*", we find no reason to change our views and stand by the original conclusions of Thomsen *et al.* (2016).

| Sample | $^{14}$C Predicted dose (Gy) | MG dose (Gy) | P-value | MG/$^{14}$C | SG CAM dose (Gy) | P-value | SG/$^{14}$C | SG/MG |
|---|---|---|---|---|---|---|---|---|
| 092201 | 103.6 ± 5.2 | 108.9 ± 2.4 | 0.50 | 1.05 ± 0.06 | 92.9 ± 2.9 | 0.07 | 0.90 ± 0.05 | 0.85 ± 0.03 |
| 092202 | 109.9 ± 5.4 | 106.4 ± 2.9 | 0.37 | 0.97 ± 0.05 | 97.3 ± 3.2 | 0.04 | 0.88 ± 0.05 | 0.91 ± 0.04 |
| 092203 | 74.3 ± 3.6 | 72.3 ± 2.1 | 0.35 | 0.97 ± 0.06 | 63.1 ± 1.9 | 0.01 | 0.85 ± 0.05 | 0.87 ± 0.04 |
| 092204 | | 69.2 ± 2.2 | | | 57.1 ± 1.9 | | | 0.83 ± 0.04 |

**Table 1. The preferred $^{14}$C ages (see page 79-80 in Thomsen *et al.*, 2016) have been converted into $^{14}$C predicted dose at 68% CI for direct comparison with both multi-grain (MG) and single-grain (SG) CAM (Central Age Model, Galbraith *et al.*, 1999) doses. A two-sample $\chi^2$ homogeneity test (e.g. Galbraith and Roberts, 2012) has been used to assess the null hypothesis that the $^{14}$C predicted dose is not statistically significantly different from the measured quartz dose. A P-value of less than 0.05 indicates that the measured dose is, by normal standards, significantly different from the independent $^{14}$C age control. Note that OSL sample 092204 has been omitted from this comparison as the independent age control is too wide to be of use.**

## REFERENCES


Arnold, L.J., Roberts, R.G., 2009. Stochastic modelling of multi-grain equivalent dose (De) distributions: implications for OSL dating of sediment mixtures. Quat. Geochronol. 4, 204-230.

Arnold, L. J., Demuro, M., 2015. Insights into TT-OSL signal stability from single-grain analysis of known age deposits at Atapuerca, Spain. Quat. Geochron. 30, 472-478.

Bailey, R.M., Arnold, L.J., 2006. Statistical modelling of single grain quartz De distributions and an assessment of procedures for estimating burial dose. Quat. Sci. Rev. 25, 2475-2502.

Costas, I., Reimann, T., Tsukamoto, S., Ludwig, J., Lindhorst, S., Frechen, M., Hass, C., Betzler, C., 2012. Comparison of OSL ages from young dune sediments with a high-resolution independent age model. Quaternary Geochronology 10, 16-23.

Delong, S.B., Arnold, L.J., 2007. Dating alluvial deposits with optically stimulated luminescence, AMS 14C and cosmogenic techniques, western Transverse Ranges, California, USA. Quaternary Geochronology 2, 129–136

Demuro, M., Arnold, L. J., Froese, D. G., Roberts, R. G., 2013. OSL dating of loess deposits bracketing Sheep Creek tephra beds, northwest Canada: dim and problematic single-grain OSL characteristics and their effect on multi-grain age estimates. Quat. Geochron. 15, 67-87.

Doerschner, N., Hernandez, M., Fitzsimmons, K.E., 2016. Sources of variability in single grain dose recovery experiments: Insights from Moroccan and Australian samples. Ancient TL, Vol. 34, No. 1, 14-25.

Douka, K., Jacobs, Z., Lane, C., Grün, R., Farr, L., Hunt, C., Inglis, R., Reyonds, T., Albert P., Aubert, M., Cullen, V., Hill, E., Kinsley, L. Roberts, R. G., Tomlinson, E. L., Wulf, S. Barker, G., 2014. The chronostratigraphy of the Haaua Fteah cave (Cyrenaica, northeast Libya). J. Hum. Evol. 66, 39-63.

Fattahi, M., Nazari, H., Bateman, M.D., Meyer, B., Sébrier, M., Talebian, M., Le Dortz, K., Foroutan, M., Ahmadi Givi, F., Ghorashi, M., 2010. Refining the OSL age of the last earthquake on the Dheshir fault, Central Iran. Quaternary Geochronology 5, 286–292.

Feathers, J. K., Piló, L., Arroyo-Kalin, M., Kipnes, R., Coblenz, D., 2010. How old is Luzia? Luminescence dating and stratigraphic integrity at Lapa Vermelha, Lagoa Santa, Brazil. Geoarchaeol. 25, 395-436.

Fuchs, M., Owen, L., 2008. Luminescence dating of glacial and associated sediments: review, recommendations and future directions. Boreas, Vol. 37, pp. 636–659.

Gaar, D., Preusser, F., 2012. Luminescence dating of mammoth remains from northern Switzerland. Quaternary Geochronology 10, 257-263

Galbraith, R., Roberts, R.G., Laslette, G., Yoshidha, Olley, J., 1999. Optical dating of single and multiple grain quartz from Jinmium Rock Shelter, Northern Australia. Part I, experimental design and statistical models. Archaeometry 41, 339-364.

Galbraith, R.F., 2005. Statistics for Fission Track Analysis. Chapman & Hall, London.

Galbraith, R.F., Roberts, R.G., 2012. Statistical aspects of equivalent dose and error calculation and display in OSL dating: an overview and some recommendations. Quat. Geochronol. 11, 1-27.

Galbraith, R.F., 2015a. On the mis-use of mathematics: A comment on “How confident are we about the chronology of the transition between Howieson's Poort and Still Bay?” by Guérin et al. (2013). Journal of Human Evolution 80, 184-186.

Galbraith, R., 2015b. A note on OSL age estimates in the presence of dose rate heterogeneity. Ancient TL, Vol. 33, No. 1, 31-34.

Geach, M. R., Thomsen, K. J., Buylaert, J.-P., Murray, A. S., Mather, A. E., Telfer, M. W., 2015. Single-grain and multi-grain OSL dating of river terrace sediments in the Tabernas Basin, SE Spain. Quat. Geochron. 2015, 213-218.

Guérin, G., Murray, A.S., Jain, M., Thomsen, K.J., Mercier, N., 2013. How confident are we in the chronology of the transition between Howieson's Poort and still Bay? J. Hum. Evol. 64, 314-317.

Guérin, G., Combès, B., Lahaye, C., Thomsen, K.J., Tribolo, C., Urbanova, P., Guibert, P., Mercier, N., Valladas, H., 2015a. Testing the accuracy of a Bayesian central-dose model for single-grain OSL, using known-age samples. Radiation Measurements 81, 62-70

Guérin, G., Frouin, M., Talamo, S., Aldeias, V., Bruxelles, L., Chiotti, L., Dibble, H.L., Goldberg, P., Hublin, J.-J., Jain, M., Lahaye, C., Madelaine, S., Maureille, B., McPherron, S.P., Mercier, N., Murray, A.S., Sandgathe, D., Steele, T.E., Thomsen, K.J., Turq, A., 2015b. A multi-method luminescence dating of the Palaeolithic sequence of La Ferrassie based on new excavations adjacent to the La Ferrassie 1 and 2 Skeletons. J. Archaeol. Sci. 58, 147-166.

Guérin, G., Frouin, M., Tuquoi, J.; Thomsen, K.J., Goldberg, P., Aldeias, V., Lahaye, Ch., Mercier, N., Guibert P., Jain, M., Sandgathe, D., McPherron, S.P., Turq, A., Dibble, H.L., 2016. The complementarity of luminescence dating methods illustrated on the Mousterian sequence of the Roc de Marsal: A series of reindeer-dominated, Quina Mousterian layers dated to MIS 3. Quaternary International, in press.

Hansen, V., Murray, A.S., Buylaert, J.-P., Yeo, E.-Y., Thomsen, K.J., 2015. A new irradiated quartz for beta source calibration. Radiat. Meas. 81, 123–127

Jacobs, Z., Duller, G.A.T., Wintle, A.G., 2006. Interpretation of single grain De distributions and calculation of De. Radiat. Meas. 41, 264-277.

Jacobs, Z., Roberts, R.G., Nespoulet, R., El Hajraoui, M.A., Debenath, A., 2012. Single grain OSL chronologies for Middle Palaeolithic deposits at El Mnasra and El Harhoura 2, Morocco: implications for Late Pleistocene huma-environment interactions along the Atlantic coast of northwest Africa. J. Hum. Evol. 62, 377-394.

Jacobs, Z., Li, B., Jankowski, N., Soressi, M., 2015. Testing of a single grain OSL chronology across the Middle to Upper Palaeolithic transition at Le Cottés (France). J. Archael. Sci. 54, 110-122.

Jacobs, Z., Roberts, R.G., 2015. An improved single grain OSL chronology for the sedimentary deposits from Diepkloof Rockshelter, Western Cape, South Africa. Journal of Archaeological Science 63, 175-192.

Kristensen, J.A., Thomsen, K.J., Murray, A.S., Buylaert, J.P., Jain, M., Breuning-Madsen, H., 2015. Quantification of termite bioturbation in a savannah ecosystem: application of OSL dating. Quat. Geochronol. 30, 334-341.

Murray, A.S., Wintle, A.G., 2003. The single aliquot regenerative dose protocol: potential for improvements in reliability. Radiat. Meas. 37, 377-381.

Olley, J.M., Deckker, P.D., Roberts, R.G., Fifield, L.K., Yoshida, H., Hancock, G., 2004b. Optical dating of deep-sea sediments using single grains of quartz: a comparison with radiocarbon. Sediment. Geol. 169, 175-189.

Reimann, T., Lindhorst, S., Thomsen, K.J., Murray, A.S., Hass, C.H., Frechen, M., 2012. OSL dating of mixed coastal sediments from sylt (German Bight, North Sea). Quaternary Geochronology 11, 52-67.

Stone, A.E.C., Thomas, D.S.G., Viles, H.A., 2010. Late Quaternary palaeohydrological changes in the northern Namib Sand Sea: New chronologies using OSL dating of interdigitated aeolian and water-lain interdune deposits. Palaeogeography, Palaeoclimatology, Palaeoecology 288, 35–53.

Thomsen, K. J., Murray, A. S., Buylaert, J. P., Jain, M., Hansen, J. H., Aubry, T., 2016. Testing single-grain quartz OSL methods using sediment samples with independent age control from the Bordes-Fitte rockshelter (Roches d'Abilly site, central France). Quat. Geochron. 31, 77-96.

Zhao, Q.Y., Thomsen, K.J., Murray, A.S., Wei, M.J., Pan, B.L., Zhou, R., Zhao, X.H., Chen, H.Y., 2015. 'Testing the use of OSL from quartz grains for dating debris flows in Miyun, northeast Beijing, China. Quaternary Geochronology 30, 320–327